Protein-protein interaction network inference with semi-supervised Output Kernel Regression

Céline BROUARD[1], Marie SZAFRANSKI[2,1] and Florence D'ALCHÉ-BUC[1,3]

[1] IBISC, EA 4526, 23 bd de France, Université d'Évry Val d'Essonne, 91037 Évry cedex, France
{celine.brouard, florence.dalche,marie.szafranski}@ibisc.fr
[2] ÉNSIIE, 1 square de la résistance, 91025 Évry cedex, France
[3] LRI, UMR CNRS 8623, bât 650, Université Paris-Sud 11, 91405 Orsay Cedex, France

Abstract *In this work, we address the problem of protein-protein interaction network inference as a semi-supervised output kernel learning problem. Using the kernel trick in the output space allows one to reduce the problem of learning from pairs to learning a single variable function with values in a Hilbert space. We turn to the Reproducing Kernel Hilbert Space theory devoted to vector- valued functions, which provides us with a general framework for output kernel regression. In this framework, we propose a novel method which allows to extend Output Kernel Regression to semi-supervised learning. We study the relevance of this approach on transductive link prediction using artificial data and a protein-protein interaction network of S. Cerevisiae using a very low percentage of labeled data.*

Keywords Protein-protein interactions, Link prediction, Kernel methods, Operator-valued kernel, RKHS, Semi-supervised learning, Transductive learning.

1 Background

Recent years have witnessed a surge of interest for network inference in biological networks. *In silico* inference of protein-protein interaction (PPI) networks is motivated by the cost and the difficulty to experimentally detect physical interactions between proteins. It mainly relies on the assumption that some input features relative to the proteins, such as amino acids sequences, gene expressions or localizations, could provide valuable information about the presence or the absence of a physical interaction. Two main approaches are devoted to PPI network inference: supervised approaches, which aim at building a pairwise classifier able to predict if two proteins interact from labeled pairs of proteins [1,2,3,4,5], and matrix completion approaches [6,7].

Let us define $\mathcal{O}$ the set of descriptions of the proteins we are interested in. Let $f : \mathcal{O} \times \mathcal{O} \to \{0, 1\}$ be a classifier that, given the predictions of two proteins, predicts if these proteins interact or not. In this work, we have chosen to convert the binary pairwise classification task into an output kernel learning task, referred as Output Kernel Regression (OKR). As in [3,4], we assume the existence of an output kernel $\kappa_y : \mathcal{O}\times\mathcal{O} \to \mathbb{R}$ that encodes the proximities of proteins in terms of nodes in the interaction network. Then, given an approximation $\widehat{\kappa_y}$ of this output kernel, we can define a classifier f_θ by thresholding its output values:

$$f_\theta(o, o') = \mathrm{sgn}(\widehat{\kappa_y}(o, o') - \theta)\,.$$

κ_y being a positive semi-definite kernel, there exists an Hilbert space $\mathcal{F}_y$, called the feature space, and a feature map $y : \mathcal{O} \to \mathcal{F}_y$ such that $\forall (o, o') \in \mathcal{O} \times \mathcal{O}, \kappa_y(o, o') = \langle y(o), y(o')\rangle_{\mathcal{F}_y}$. In output kernel regression, we build an approximation of κ_y from the inner product between the outputs of a single input function $h : \mathcal{O} \to \mathcal{F}_y : \widehat{\kappa_y}(o, o') = \langle h(o), h(o')\rangle_{\mathcal{F}_y}$. Using the kernel trick in the output space allows one to reduce the problem of learning from pairs to learning a single variable function h with values in a Hilbert space (the output feature space $\mathcal{F}_y$).

Previous works developed tree-based Output Kernel Regression models by extending multiple output regression trees (OK3) and ensemble methods to output feature space endowed with a kernel [3,4]. In this work, we propose a novel method, which allows to extend Output Kernel Regression to the semi-supervised framework.

2 Semi-supervised Output Kernel Regression

In biology, it is often the case that labeled pairs of proteins are difficult to obtain, due to the cost and the time needed for experimental methods. However, additional description of protein properties are often available. This motivates for dealing with the problem using semi-supervised approaches. Graph-based regularization is a powerful approach to semi-supervised regression that enforces the smoothness of the function, permitting to propagate output labels over close inputs [11,12]. Belkin et al. [12] have proposed to explicitly embed such ideas into the framework of regularization within Reproducing Kernel Hilbert Space (RKHS) for real-valued functions.

In the context of OKR, the function to be learnt is not real-valued but vector-valued in the output Hilbert space. If we want to take benefit from the theoretical framework of Reproducing Hilbert Space theory (RKHS), well appropriate for regularization, we need to turn to the proper RKHS theory, devoted to vector-valued functions [9,10]. This theory requires to define an operator-valued kernel instead of a scalar input kernel. As in RKHS theory with scalar valued functions, representer theorems for different loss functions can be proven.

Let $\{(o_i, \mathbf{y}_i)\}_{i=1}^{\ell}$ be a set of labeled examples and $\{o_i\}_{i=\ell+1}^{\ell+u}$ a set of unlabeled examples. Let $\mathcal{H}$ be a RKHS with reproducing kernel $\mathcal{K}_x$, and a symmetric matrix W with positive values measuring the similarity of proteins in the input space. In this work, we propose to learn the vector-valued function h by minimizing a penalized least square cost function with a smoothness constraint :

$$\underset{h\in\mathcal{H}}{\operatorname{argmin}}\, J(h) = \sum_{i=1}^{\ell} \|h(o_i) - \mathbf{y}_i\|^2_{\mathcal{F}_y} + \lambda_1 \|h\|^2_{\mathcal{H}} + \lambda_2 \sum_{i,j=1}^{\ell+u} W_{ij} \|h(o_i) - h(o_j)\|^2_{\mathcal{F}_y} \,, \tag{1}$$

with λ_1 and $\lambda_2 > 0$.

We stated and proved a new representer theorem devoted to semi-supervised learning with a penalized least-square cost. Then, given a simple definition of the operator-valued kernel based on some input scalar kernel, we derived a close-formed solution that extends the reformulated KDE proposed by Cortes et al. [14] to the semi-supervised case. [1]

3 Experiments

3.1 Experimental protocol

The approach was evaluated in the transductive setting. For different percentage values of labeled proteins, we randomly picked a subsample of proteins as labeled examples. Labeled interactions correspond to interactions between two labeled proteins, and we searched to predict the remaining interactions. A 10% selection of labeled proteins therefore corresponds to only 1% labeled interactions. We evaluated the performance by averaging the AUC-ROC over 10 random choices of the training set.

3.2 Results

We extensively studied the behaviour of the provided model using an artificial dataset produced by sampling random graphs form a Erdős-Renyi law with different probabilities of presence of edges. The input features were obtained by applying Kernel Principal Component Analysis on the diffusion kernel associated with the graph, and using the components capturing 95% of the variance. We observe from the results obtained that the semi-supervised approach improves upon the supervised one on AUC-ROC, especially for a small percentage of labeled data (up to 10%). Based on these results one can formulate the hypothesis that supervised link prediction is harder in the case of more dense networks and that the contribution of unlabeled data seems more helpful in this case. One can also assume that using unlabeled data increases the AUCs for low percentage of labeled data. But when enough information can be found in the labeled data, semi-supervised learning does not improve the performance.

1. The details of this method are given in the publication [13].

We also illustrated our method on a protein-protein interaction network of the yeast *S. Cerevisiae* composed of 984 proteins linked by 2438 interactions. To reconstruct the PPI network, we used gene expression data, phylogenetic profiles, protein localization and protein interaction data derived from yeast two-hybrid experiments as input features [2,3,4,5,6]. For each of these features, our method compares favorably with the existing methods in the supervised setting [13]. In a second step, we experimented the method in the transductive setting using gene expression as input features. Fig. 1 reports the averaged AUC-ROC and the standard deviations for different values of λ_2 and different percentages of labeled proteins. One can see that the semi-supervised approach improves the AUC-ROC upon the supervised one (corresponding to $\lambda_2 = 0$) for all percentage values. This improvement is especially significant when the percentage of labeled proteins is low, which is usually the case in PPI network inference problems.

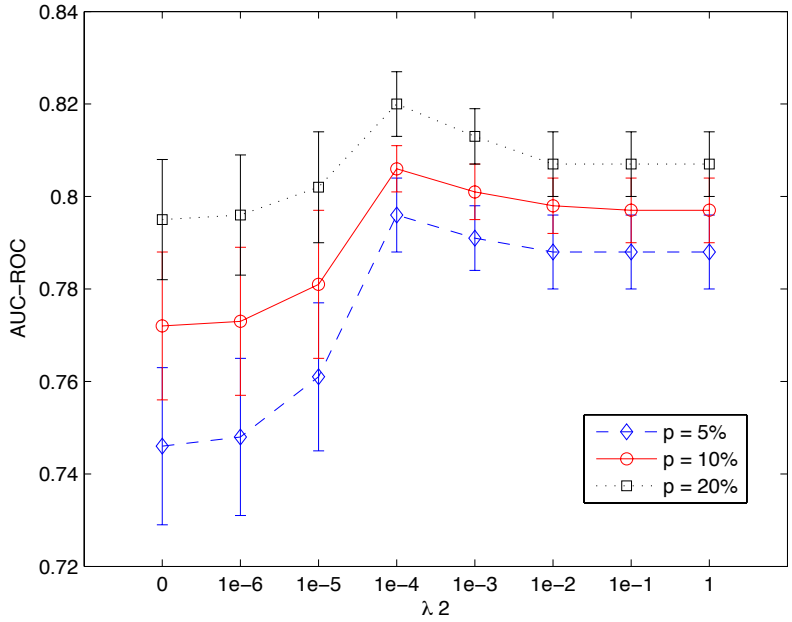

Figure 1. Averaged AUC-ROC results for the reconstruction of the Yeast PPI network from gene expression data in the supervised and semi-supervised settings. The percentage values correspond to the proportions of labeled proteins.

References

[1] A. Ben-Hur and W.S. Noble, Kernel methods for predicting protein-protein interactions, *Bioinformatics*, vol. 21, pp. 38-46, 2005.

[2] Y. Yamanishi, J.-P. Vert and M. Kanehisa, Protein network inference from multiple genomic data: a supervised approach, *Bioinformatics*, vol. 20, pp. 363-370, 2004.

[3] P. Geurts, L. Wehenkel and F. d'Alché-Buc, Kernelizing the output of tree-based methods, in *Proceedings of the 23th International Conference on Machine learning*, 2006.

[4] P. Geurts, N. Touleimat, M. Dutreix and F. d'Alché-Buc, Inferring biological networks with output kernel trees, *BMC Bioinformatics*, 8, 2007.

[5] K. Bleakley, G. Biau and J.-P. Vert, Supervised reconstruction of biological networks with local models, *Bioinformatics*, vol. 23, pp. i57-i65, 2007.

[6] T. Kato, K. Tsuda and K. Asai, Selective integration of multiple biological data for supervised network inference, *Bioinformatics*, vol. 21, pp. 2488-2495, 2005.

[7] K. Tsuda and W.S. Noble, Learning kernels from biological networks by maximizing entropy, *Bioinformatics*, vol. 20, pp. 326-333, 2004.

[8] R.I. Kondor and J.D. Lafferty, Diffusion Kernels on Graphs and Other Discrete Input Spaces, In *Proceedings of the 19th International Conference on Machine Learning*, 2002.

[9] E. Senkene and A. Tempel'man, Hilbert Spaces of operator-valued functions, *Lithuanian Mathematical Journal*, 13, pp. 665-670, 1973.

[10] C. A. Micchelli and M. A. Pontil, On Learning Vector-Valued Functions, *Neural Computation*, 17, pp. 177-204, 2005.

[11] D. Zhou, O. Bousquet, T.N. Lal, J. Weston and B. Scholkopf, Learning with Local and Global Consistency, in *Advances in Neural Information Processing Systems 16*, 2004.

[12] M. Belkin, P. Niyogi and V. Sindhwani, Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples, *Journal of Machine Learning Research*, 7, pp. 2399-2434, 2006.

[13] C. Brouard, F. d'Alché-Buc and M. Szafranski, Semi-supervised Penalized Output Kernel Regression for link prediction, *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[14] C. Cortes, M. Mohri and J. Weston, A general regression technique for learning transductions, in *Proceedings of the 22nd International Conference on Machine Learning*, pp. 153-160, 2005.